

# Pumping lemma and Ogden lemma for tree-adjoining grammars

Alexey Sorokin<sup>1,2</sup>

<sup>1</sup> Moscow State University, Faculty of Mathematics and Mechanics

<sup>2</sup> Moscow Institute of Physics and Technology,  
Faculty of Innovations and High Technologies

**Abstract.** The pumping lemma and Ogden lemma offer a powerful method to prove that a particular language is not context-free. In 2008 Kanazawa proved an analogue of pumping lemma for well-nested multiple-context free languages. However, the statement of lemma is too weak for practical usage. We prove a stronger variant of pumping lemma and an analogue of Ogden lemma for the class of tree-adjoining languages. We also use these statements to prove that some natural context-sensitive languages cannot be generated by tree-adjoining grammars.

## 1 Introduction

It is known since the 80-s that context-free grammars are not suitable for proper analyse and description of natural language syntax. The class of mildly context-sensitive languages ([1] was an informal attempt to capture the degree of context-sensitivity required for most common language phenomena keeping as much advantages of context-free grammars as possible. The most desirable properties to preserve are the feasible polynomial complexity of parsing, the independence of derivation from the context (the notion of context had to be extended to capture long-distance dependencies) and the existence of convenient normal forms.

Well-nested multiple context-free languages (wMCFLs) is between the candidates to satisfy these requirements, see [4] for discussion. The corresponding grammar formalism, well-nested MCFG or wMCFG, is defined as a subclass of multiple context-free grammars (MCFGs, [8]) with rules of special form providing the correct embedding of constituents. In particular, 2-wMCFGs are equivalent to tree-adjoining grammars ([11], [2]) and then to head grammars ([7]).

We find it sensible to think of wMCFGs not as the restriction of MCFGs, but as the generalization of head grammars. Our approach bases on two principal ideas. The first is to derive not words but terms whose values are the words of the language. Then the generative power of a grammar formalism essentially depends on the set of term connectives and their interpretation as language operations. For example, in the case of context-free grammars the set of connectives includes concatenation only. Due to its associativity, the terms are just words consisting of terminals and nonterminals. Though our approach seems to be redundant there, we find it useful for more complex cases. Its principal advantage is that the structure of derivation trees remains “context-free” for any set of connectives.

To simulate well-nested MCFGs we use the intercalation connectives. The binary operation  $\odot_j$  of  $j$ -intercalation replaces the  $j$ -th separator in its first argument by its second argument (for example,  $a1b1c \odot_2 a1b = a1ba1bc$ ). It is straightforward to prove that all “well-nested” combinations of constituents can be presented using only intercalation and concatenation operations.

The exact generative power of wMCFGs is not known. Moreover, some languages are supposed to be not wMCFLs, although they are not proved to be outside this family. The most known example is the MIX language  $\{w \in \{a, b, c\}^{\mathbb{N}} \mid |w|_a = |w|_b = |w|_c\}$ . It was shown in [4] to be not a 2-MCFL, but the proof used combinatorial and geometric arguments which are troublesome to be generalized for the class of all wMCFGs. Though Kanazawa presented an analogue of pumping lemma for wMCFLs in [3], this lemma is too weak to be applied in the general case, since it does not impose any conditions on the length and position of the pumped segment. We prove a stronger version of the pumping lemma and an analogue of Ogden lemma ([6]) for tree-adjointing grammars basing on the ideas already used in [3]. Our variant of Ogden lemma allows us to give a simple proof of the fact that MIX cannot be generated by a tree-adjointing grammar.

We suppose the reader to be familiar with the basics of formal languages theory nevertheless all the required definitions are explicitly formulated.

## 2 Preliminaries

### 2.1 Terms and their equivalence

In this section we define displacement context-free grammars (DCFGs) which are a more “purely logical” reformulation of well-nested MCFGs. The first subsection is devoted to the notions of term and context which play the key role in the architecture of DCFGs, it also contains some results on term equivalence which are extensively used in the further. We mostly follow the definitions from [9].

Let  $\Sigma$  be a finite alphabet and 1 be a distinguished separator,  $1 \notin \Sigma$ . For every word  $w \in (\Sigma \cup 1)^*$  we define its rank  $rk(w) = |w|_1$ . We define the  $j$ -th intercalation operation  $\odot_j$  which consists in replacing the  $j$ -th separator in its first argument by its second argument. For example,  $a1b11d \odot_2 c1c = a1bc1c1d$ .

Let  $k$  be a natural number and  $N$  be the set of nonterminals. The function  $rk: N \rightarrow \overline{0, k}$  assigns every element of  $N$  its rank. Let  $Op_k = \{., \odot_1, \dots, \odot_k\}$  be the set of binary operation symbols, then the ranked set of  $k$ -correct terms  $Tm_k(N, \Sigma)$  is defined in the following way (we write simply  $Tm_k$  in the further):

1.  $N \subset Tm_k(N, \Sigma)$ ,
2.  $\Sigma^* \subset Tm_k(N, \Sigma)$ ,  $\forall w \in \Sigma^* rk(w) = 0$ ,
3.  $1 \in Tm_k$ ,  $rk(1) = 1$ ,
4. If  $\alpha, \beta \in Tm_k$  and  $rk(\alpha) + rk(\beta) \leq k$ , then  $(\alpha \cdot \beta) \in Tm_k$ ,  
 $rk(\alpha \cdot \beta) = rk(\alpha) + rk(\beta)$ .
5. If  $j \leq k$ ,  $\alpha, \beta \in Tm_k$ ,  $rk(\alpha) + rk(\beta) \leq k + 1$ ,  $rk(\alpha) \geq j$ , then  
 $(\alpha \odot_j \beta) \in Tm_k$ ,  $rk(\alpha \odot_j \beta) = rk(\alpha) + rk(\beta) - 1$ .

We refer to the elements of the set  $N \cup \Sigma^* \cup \{1\}$  as basic subterms. We will often omit the symbol of concatenation and assume that concatenation has greater priority than intercalation, so  $Ab \odot_2 cD$  means  $(A \cdot b) \odot_2 (c \cdot D)$ . This simplification allows us to consider words in the alphabet  $\Sigma_1^*$  as terms either. The set of  $k$ -correct terms includes all the terms of sort  $k$  or less that also do not contain subterms of rank greater than  $k$ .

Let  $\nu$  be a valuation function assigning every element of  $N$  a subset of  $\Sigma_1^*$ , such that  $\nu(A)$  includes only the words with rank  $\text{rk}(A)$ . Then every term  $\alpha$  is mapped to some language of words with rank  $\text{rk}(\alpha)$ , we call this language the value of  $\alpha$ . A term is ground if it does not contain nonterminals, the value of a ground term is the same for all valuations. Two terms  $\alpha$  and  $\beta$  are equivalent if their values are equal for any valuation, the equivalence relation is denoted by  $\sim$ . Note that every ground term  $\gamma$  is equivalent to its value  $\nu(\gamma)$ . The next lemma follows from the definitions.

**Lemma 1.** *The following equivalences hold for any terms  $\alpha, \beta, \gamma$  such that the terms in lemma are correct:*

1.  $(\alpha \cdot \beta) \cdot \gamma \sim \alpha \cdot (\beta \cdot \gamma)$ ,
2.  $(\alpha \cdot \beta) \odot_j \gamma \sim (\alpha \odot_j \gamma) \cdot \beta$  if  $j \leq \text{rk}(\alpha)$ ,
3.  $(\alpha \cdot \beta) \odot_j \gamma \sim \alpha \cdot (\beta \odot_{j-\text{rk}(\alpha)} \gamma)$  if  $\text{rk}(\alpha) < j \leq \text{rk}(\alpha) + \text{rk}(\beta)$ ,
4.  $(\alpha \odot_l \beta) \odot_j \gamma \sim (\alpha \odot_j \gamma) \odot_{l+\text{rk}(\gamma)-1} \beta$  if  $j < l$ ,
5.  $(\alpha \odot_l \beta) \odot_j \gamma \sim \alpha \odot_l (\beta \odot_{j-l+1} \gamma) \odot \beta$  if  $l \leq j < l + \text{rk}(\beta)$ ,
6.  $(\alpha \odot_l \beta) \odot_j \gamma \sim (\alpha \odot_{j-\text{rk}(\beta)+1} \gamma) \odot_l \beta$  if  $j \geq l + \text{rk}(\beta)$ .
7.  $1 \odot_1 \alpha \sim \alpha$ ,
8.  $\alpha \odot_j 1 \sim \alpha$  for any  $j \leq \text{rk}(\alpha)$ .

A context  $C[]$  is a term with a distinguished placeholder  $\#$  instead one of its leafs. If  $\beta$  is a term, then  $C[\beta]$  denotes the result of replacing  $\#$  by  $\beta$  (provided the created term is correct). For example,  $C[] = b1 \odot_1 (a \cdot \#)$  is a context and  $C[A \cdot c] = b1 \odot_1 aAc$ . Note that if the terms  $\alpha$  and  $\beta$  are equivalent then the equivalence  $C[\alpha] \sim C[\beta]$  holds for any context  $C$ .

With every term  $\alpha$  we associate its syntactic tree  $\text{tree}(\alpha)$  in a natural way. Then subterms of  $\alpha$  correspond to the nodes of this tree and vice versa, a subterm is internal if it corresponds to an internal node (it means the subterms contains a binary connective). A term  $k$ -essential if its rank is less than  $k$ , as well as the rank of its atomic subterms. The next lemma is proved in the Appendix A.

**Lemma 2.** *For any  $k$ -essential term  $\alpha$  there is an equivalent  $k$ -correct term  $\alpha'$ .*

## 2.2 Displacement context-free grammars

This subsection introduces the notion of a displacement context-free grammar.

**Definition 1.** *A  $k$ -displacement context-free grammar ( $k$ -DCFG) is a quadruple  $G = \langle N, \Sigma, P, S \rangle$ , where  $\Sigma$  is a finite alphabet,  $N$  is a finite ranked set of nonterminals and  $\Sigma \cap N = \emptyset$ ,  $S \in N$  is a start symbol such that  $\text{rk}(S) = 0$  and  $P$  is a set of rules of the form  $A \rightarrow \alpha$ . Here  $A$  is a nonterminal,  $\alpha$  is a term from  $\text{Term}_k(N, \Sigma)$ , such that  $\text{rk}(A) = \text{rk}(\alpha)$ .*

**Definition 2.** The derivability relation  $\vdash_G \in N \times Tm_k$  associated with the grammar  $G$  is the smallest reflexive transitive relation such that the facts  $(B \rightarrow \beta) \in P$  and  $A \vdash C[B]$  imply that  $A \vdash C[\beta]$  for any context  $C$ . If the set of words derivable from  $A \in N$  is  $L_G(A) = \{\nu(\alpha) \mid A \vdash_G \alpha, \alpha \in GrTm_k\}$ , then  $L(G) = L_G(S)$ .

*Example 1.* Let the  $i$ -DCFG  $G_i$  be the grammar  $G_i = \langle \{S, T\}, \{a, b\}, P_i, S \rangle$ . Here  $P_i$  is the following set of rules (notation  $A \rightarrow \alpha | \beta$  means  $A \rightarrow \alpha, A \rightarrow \beta$ ):

$$\begin{aligned} S &\rightarrow \underbrace{(\dots (aT \odot_1 a) + \dots)}_{i-1 \text{ times}} \odot_1 a \mid \underbrace{(\dots (bT \odot_1 b) + \dots)}_{i-1 \text{ times}} \odot_1 b \\ T &\rightarrow \underbrace{(\dots (aT \odot_1 1a) + \dots)}_{i-1 \text{ times}} \odot_i 1a \mid \underbrace{(\dots (bT \odot_1 1b) + \dots)}_{i-1 \text{ times}} \odot_i 1b \mid 1^i \end{aligned}$$

$G_i$  generates the language  $\{w^{i+1} \mid w \in \{a, b\}^+\}$ . For example, this is the derivation of the word  $(aba)^3$  in  $G_2$ :  $S \rightarrow (aT \odot_1 a) \odot_1 a \rightarrow (a((bT \odot_1 1b) \odot_2 1b) \odot_1 a) \odot_1 a \rightarrow (a((b((aT \odot_1 1a) \odot_2 1a) \odot_1 1b) \odot_2 1b) \odot_1 a) \odot_1 a \rightarrow (a((b((a11 \odot_1 1a) \odot_2 1a) \odot_1 1b) \odot_2 1b) \odot_1 a) \odot_1 a = (a(b(a1a1a \odot_1 1b) +_2 1b) +_1 a) \odot_1 a = (aba1ba1ba \odot_1 a) \odot_1 a = abaabaaba$ .

Two  $k$ -DCFGs are equivalent if they generate the same language. Since internal nodes of terms in a  $k$ -DCFG rules are also of rank  $k$  or less, the  $k$ -DCFGs can be binarized just like the context-free grammars to obtain a variant of Chomsky normal form. Precisely, the following theorem holds (see [9] for details):

**Theorem 1.** Every  $k$ -DCFG is equivalent to some  $k$ -DCFG  $G = \langle N, \Sigma, P, S \rangle$  which has the rules only of the following form:

1.  $A \rightarrow B \cdot C$ , where  $A \in N$ ,  $B, C \in N - \{S\}$ ,
2.  $A \rightarrow B \odot_j C$ , where  $j \leq k$ ,  $A \in N$ ,  $B, C \in N - \{S\}$ ,
3.  $A \rightarrow a$ , where  $A \in N$ ,  $a \in \Sigma_1$ ,
4.  $S \rightarrow \epsilon$ .

We have already mentioned that  $k$ -DCFGs are equivalent to  $(k+1)$ -wMCFGs. In the case of  $k = 1$  this statement is straightforward since both 1-DCFGs and 2-wMCFGs are just reformulations of Pollard wrap grammars ([7]). We will not recall the definitions of a wMCFG, the interested reader may consult [8] and [3].

### 3 Terms and derivations in DCFGs

In this section we investigate more thoroughly the properties of terms and derivation in DCFGs. At first we give some fundamental notions. We assume that all the grammars are in Chomsky normal form.

**Definition 3.** A node  $v'$  in the syntactic tree is a direct descendant of a node  $v$  if  $rk(v') = rk(v)$ ,  $v'$  is a descendant of  $v$  and all the nodes on the path from  $v$  to  $v'$  has rank the same rank as  $v$  and  $v'$ . A subterm  $\beta$  is a direct subterm of a term  $\alpha$ , if its root node is the direct descendant of the root of  $\alpha$ .

**Definition 4.** A context  $C$  is called *ground* if it does not contain nonterminals.

Let  $\alpha$  be a term of rank  $l$ , we denote by  $\alpha \otimes (u_1, \dots, u_l)$  the result of simultaneous replacement of all the separators in  $\alpha$  by  $u_1, \dots, u_l$ .

**Lemma 3.** Let  $\alpha = C[\beta]$  for some ground context  $C$  and term  $\beta$  of rank  $l$ . There exist words  $x_1, x_2, u_1, \dots, u_l \in \Sigma_1^*$  depending only from the context  $C$  such that  $\alpha \sim x_1(\beta \otimes (u_1 \dots u_l))x_2$  and  $\text{rk}(\alpha) = \text{rk}(x_1) + \text{rk}(x_2) + \sum_{i=1}^l \text{rk}(u_i)$ .

*Proof.* Induction on the structure of the context  $C$ . The induction step uses Lemma 1 and the equivalence between a ground term and its value.

**Lemma 4.** Let  $\beta$  be a direct subterm of a term  $\alpha$  and  $C$  be the ground context such that  $\alpha = C[\beta]$ . Then the equivalence  $\alpha \sim x_1\beta \otimes (y_1 1_{z_1}, \dots, y_l 1_{z_l})x_2$  holds for some words  $x_1, x_2, y_1, z_1, \dots, y_l, z_l \in \Sigma^*$ , depending only from the context  $C$ .

*Proof.* Induction on the structure of the context  $C$ , the base is trivial. On the induction step consider the root connective of term  $\alpha$ . If this connective is  $\cdot$ , then  $\alpha$  has the form  $\alpha' \cdot \eta$  or  $\eta \cdot \alpha'$  for some ground term  $\eta$  of sort 0 and some term  $\alpha$  such that  $\beta$  is its direct subterm. The statement follows from the induction hypothesis with the help of the fact that  $\eta$  is equivalent to the word  $\nu(\eta) \in \Sigma^*$ .

If the root connective is  $\odot$ , then  $\alpha = \alpha' \odot_j \eta$  or  $\alpha = \eta \odot_1 \alpha'$  for some ground term  $\eta$  of sort 1 and  $\alpha'$  having a direct subterm  $\beta$ . Then the statement also easily follows from the induction hypothesis.

**Definition 5.** A  $(l, n)$ -matreshka is a descending path in a syntactic tree containing nodes  $v_1, \dots, v_n$  of rank  $l$  such that  $v_n$  is the direct descendant of  $v_1$ .

Let  $D$  be the derivation of  $\alpha$  from some nonterminal  $A$  of the grammar  $G$  (we denote it by  $D: A \vdash \alpha$ ). We associate with  $D$  its derivation tree  $T_D$  obtained by attaching nonterminals to the nodes of  $\text{tree}(\alpha)$ . The labeling procedure is the following: if the last step of  $D$  applied the rule  $B \rightarrow \beta$  in the context  $C$  then we label by  $B$  the root node of the inserted subtree and keep other labels unchanged. Since  $G$  is in Chomsky normal form, only the nonterminal leaves of  $\text{tree}(\alpha)$  are unlabeled. Then we label every such node by the nonterminal it contains. The lemma below is proved by induction on derivation length.

**Lemma 5.** Let  $A \vdash \alpha$  and  $T_D$  be the corresponding derivation tree. For every representation  $\alpha = C[\beta]$  there are derivations  $D_1: A \vdash C[B]$  and  $D_2: B \vdash \beta$  such that  $T_D$  is obtained by replacing  $B$  with  $D_2$  in the context  $C$ .

A rule  $A \rightarrow \alpha$  is derivable in a grammar  $G$  if  $A \vdash_G \alpha$ . Adding derivable rules to a grammar does not change the language it generates. Rules  $A \rightarrow \alpha$  and  $A \rightarrow \alpha'$  are called equivalent if the terms  $\alpha$  and  $\alpha'$  are equivalent. If one of such rules is already in  $G$ , adding the other does not affect the generated language.

Since this moment we focus our attention on 1-DCFGs only.

**Definition 6.** We call a term  $\alpha \in Tm_1$  1-internal if all its subterms are of rank 1 possibly except the term itself and its nonterminal leaves.

**Definition 7.** We call a 1-DCFG  $G$   $t$ -duplicated if for every derivable rule  $A \rightarrow \alpha$ , such that  $\text{rk}(A) = 0$ ,  $d(\alpha) \leq t$  and  $\alpha$  is a 1-internal term, there is another derivable rule  $A \rightarrow \alpha'$  such that  $\alpha' \in Tm_0$ .

The next lemma is the key technical tool for the further.

**Lemma 6.** For any 1-DCFG  $G = \langle N, \Sigma, P, S \rangle$  in Chomsky normal form and any  $t$  there is an equivalent  $t$ -duplicated 1-DCFG  $G_t = \langle N', \Sigma, P', S \rangle$  in Chomsky normal form such that the sets  $N$  and  $N'$  have common nonterminals of rank 1.

*Proof.* Let the set  $P_D$  contain all derivable rules  $A \rightarrow \alpha$  with  $\text{rk}(A) = 0$ ,  $d(\alpha) \leq t$  and 1-internal term  $\alpha$  that do not have an equivalent rule  $A \rightarrow \alpha'$  with  $\alpha' \in Tm_0$  (we call such rules unduplicated). Note that this set is finite. We proceed by induction on the cardinality of  $P_D$ , if this set is empty, then  $G$  is already  $t$ -duplicated and there is nothing to prove.

Let the rule  $A \rightarrow \alpha$  be an element of the set  $P_D$ . By Lemma 1 there is a term  $\alpha' \in Tm_0$  such that  $\alpha \sim \alpha'$ . Binarizing the rule  $A \rightarrow \alpha'$  with the help of new nonterminals and adding all the generated rules to the set  $P$  leads to an equivalent grammar  $G'$  in Chomsky normal form. Note that this operation does not require adding nonterminals of rank 1 by definition of the set  $Tm_0$ .

Let us prove there are no new rules in  $P_D$ . Indeed, if a derivable rule  $B \rightarrow \beta$  of required form has been created, then its derivation uses some rules obtained during binarization of the rule  $A \rightarrow \alpha'$ . But all such rules contain only nonterminals of rank 0. It means that  $\text{tree}(\beta)$  contains some internal node of rank 0, hence  $\beta$  is not 1-internal, which gives the contradiction.

We have duplicated rule  $A \rightarrow \alpha$ , so the cardinality of  $P_D$  has decreased and we may use the induction hypothesis. Since no nonterminals of rank 1 were added, the second statement of the lemma is also satisfied, which was required.

We call a term  $\alpha$  derivable in the grammar  $G$  if  $A \vdash \alpha$  for some nonterminal  $A$ . We call a term  $S$ -derivable, if it is derived from initial nonterminal. A term is 1-internal if it does not contain internal nodes of rank 0 possibly except the root. A term is  $t$ -correct if there are no 1-internal subtrees of depth  $t$  or less in its syntactic tree (the leaves of the subtree may be not the leaves of the whole tree). The lemmas below translate Lemma 6 to the language of derivable terms.

**Lemma 7.** If the grammar  $G$  is  $t$ -duplicated, then for every word  $w \in L(G)$  there is a  $t$ -correct term  $\alpha$  such that  $w = \nu(\alpha)$ .

**Lemma 8.** For every 1-DCFG  $G$  there is an equivalent 1-DCFG in Chomsky normal form such that for every word  $w \in L(G)$  there exists a  $t$ -correct term  $\alpha$  such that  $w = \nu(\alpha)$ .

## 4 Main contruction

In this section we deal a bit more with derivation trees to obtain the main results of the paper. If  $\alpha$  is a derivable term and  $v$  is the node of rank 0 in its

syntactic tree, then the downcut of  $v$  is a subtree  $T'$  with a root  $v$  obtained by the following procedure: on every branch descending from  $v$  we include to  $T'$  all the nodes until the node of rank 0 or the leave node is reached. In this case we also include this node to  $T'$  and turn to another branch. We call a downcut degenerate if its only internal node is its root.

**Lemma 9.** *Let  $v$  be a node of rank 0 of the syntactic tree of a  $t$ -correct term  $\alpha$ . If its downcut  $T'$  is non-degenerate, then it contains a  $(1, t)$ -matreshka including one of the children of  $v$ .*

*Proof.* Since  $T'$  is non-degenerate, it is a 1-internal subtree of  $tree(\alpha)$ . The depth of  $T'$  must be at least  $t+1$  because  $\alpha$  is  $t$ -correct. Recall that all internal nodes of a branch in 1-internal tree are of rank 1. Then the longest branch of  $T'$  contains at least  $t$  successive 1-ranked nodes which is a  $(1, t)$ -matreshka.

Let  $v$  be a node of a syntactic tree, we call the ceiling of  $v$  its closest ancestor of rank 0. If such a node  $v'$  exists, the ceiling distance of  $v$  is the length of the path from  $v$  to  $v'$ . We call the  $t$ -distance of  $v$  the number  $d_t(v)$  of edges between  $v$  and its closest ancestor (including  $v$  itself) which is an element of some  $(1, t)$ -matreshka.

**Lemma 10.** *Let  $\alpha$  be a  $t$ -correct term of rank 0, then  $d_t(v) \leq t-1$  for every node  $v$  of rank 1 in  $tree(\alpha)$ .*

*Proof.* Consider an arbitrary node  $v$  of rank 1. If its ceiling distance is greater than  $t$  then at least  $t-1$  closest ancestors of  $v$  are of rank 1 and so  $v$  is a member of a  $(1, t)$ -matreshka. Otherwise let  $v'$  be a ceiling node of  $v$ , then  $v$  is a direct descendant of its child  $v''$  of rank 1. Since  $v$  cannot have two children of rank 1,  $v''$  is a member of a  $(1, t)$ -matreshka by Lemma 9. It remains to mention that there are at most  $t-1$  edges between  $v''$  and  $v$ .

Now we are ready to prove the first principal result of our paper. The following definition is a variant of the definition of a  $k$ -pump from [3]:

**Definition 8.** *We call a  $2l$ -pump a pair of internal nodes  $v$  and  $v'$  of a derivation tree such that  $v$  and  $v'$  has the same label of rank  $l-1$  and  $v'$  is the direct descendant of  $v$ . In this case  $v$  is the top and  $v'$  — the bottom node of the pump.*

**Theorem 2.** *For any 1-DCFL  $L$  there is number  $n$ , such that any word  $w \in L$  with  $|w| > n$  can be represented in the form  $w = x_0 y_1 u_1 z_1 x_1 y_2 u_2 z_2 x_2$ , such that:*

1.  $|y_1 z_1 y_2 z_2| > 0$ ,
2.  $|y_1 u_1 z_1 y_2 u_2 z_2| \leq n$ ,
3. For any  $m \in \mathbb{N}$  the word  $x_0 y_1^m u_1 z_1^m x_1 y_2^m u_2 z_2^m x_2$  belongs to  $L(G)$ .

*Proof.* We assume that  $L$  is generated by a 1-DCFG  $G$  in Chomsky normal form with  $N_1$  nonterminals of rank 1. Let  $G'$  be an equivalent  $(N_1 + 1)$ -duplicated DCFG in Chomsky normal form, which exists by Lemma 6. Then  $G'$  has the

same number of nonterminals of rank 1 as  $G$ . Let  $G'$  contain  $N_0$  nonterminals of rank 0. We set  $N = 2N_1 + N_0$  and  $n = 2^N$ .

Let  $|w| > n$  and  $\alpha$  be  $(N_1 + 1)$ -correct  $S$ -derivable term such that  $\nu(\alpha) = w$ . Let  $D$  be the corresponding derivation tree, then the depth  $D$  is at least  $N + 1$ , we consider its longest branch. If  $N_0 + 1$  deepest nodes of it have rank 0, then by the pigeon-hole principle some node label is repeated. If conversely, a node  $v_1$  has rank 1, by Lemma 10 one of  $N_1$  closest ancestors of  $v_1$  belongs to a  $(1, N_1 + 1)$ -matreshka. Again by pigeon-hole principle some label appears twice in this matreshka. In both the cases there is a pump with the top node  $v$  and the bottom node  $v'$  such that the depth of subtree below  $v$  is not greater than  $N_0 + 2N_1 = N$ .

If it is a 2-pump, then the proof repeats the standard one used for context-free grammars. Assume that it is a 4-pump, so  $v$  and  $v'$  are of rank 1. Let  $C_1$  and  $C_2$  be the ground contexts corresponding to  $v$  and  $v'$ , then the following conditions hold for some term  $\beta$  and nonterminal  $B$  of rank 1:

1.  $\alpha = C_1[C_2[\beta]]$ ,
2.  $S \vdash C_1[B]$ ,
3.  $B \vdash C_2[B]$ ,
4.  $B \vdash \beta$ .

Let  $\nu(\beta) = u_1 u_2$ . By Lemma 4 the context  $C_2[\gamma]$  is equivalent to  $y_1(\gamma \odot_1 (z_1 1 z_2)) y_2$  for some words  $y_1, y_2, z_1, z_2 \in \Sigma^*$  for any term  $\gamma$  of rank 1. Also  $C_1[\eta] \sim x_0(\eta \odot_1 x_1) x_2$  for some words  $x_0, x_1, x_2 \in \Sigma^*$ . Then  $w$  is equivalent and hence equal to the word  $x_0((y_1((u_1 u_2) \odot_1 (z_1 1 z_2)) y_2) \odot_1 x_1) x_2 = x_0 y_1 u_1 z_1 x_1 y_2 u_2 z_2 x_2$ . The depth of  $C_2[\beta]$  is not greater than  $N$ , so its value  $y_1 u_1 z_1 y_2 u_2 z_2$  cannot be longer than  $n$ . It remains to prove the third statement.

We denote by  $C_2^m$  the context  $C_2 \underbrace{[\dots [ C_2 ] \dots]}_{(m-1) \text{ times}}$ . Repeating the derivation  $B \vdash C_2[B]$   $m$  times, we obtain the derivation  $B \vdash C_2^m[B]$ . Applying Lemma 4 to the context  $C_2$  several times and using Lemma 1, we get the equivalence  $C_2^m[\gamma] \sim y_1^m(\gamma \odot_1 (z_1^m 1 z_2^m)) y_2^m$ . Setting  $\gamma = \beta$  yields that  $y_1^m u_1 z_1^m 1 z_2^m u_2 y_2^m \in L_{G'}(B)$  and consequently  $x_0 y_1^m u_1 z_1^m x_1 z_2^m u_2 y_2^m x_2 \in L_{G'}(S)$ . The theorem is proved.

Let the pair of nodes  $v$  and  $v'$  be a  $2l$ -pump. We call its collapsing the replacement of subtree growing from  $v$  by subtree growing from  $v'$ . The scope of a  $2l$ -pump consists of the nodes being descendants of  $v$  but not of  $v'$ ; all such nodes are removed when collapsing this pump.

**Lemma 11.** *Let  $T'$  be a tree obtained from  $T$  by collapsing some pump. If the nodes  $v_1$  and  $v_2$  form a pump in  $T'$ , then they have also formed a pump in  $T$ .*

*Proof.* Let  $v$  and  $v'$  be, respectively, the top and bottom nodes of the collapsed pair. If  $v'$  is not on the path from  $v_1$  to  $v_2$  in  $T'$  then  $v_2$  has already been a direct descendant of  $v_1$  in  $T$ . Otherwise  $(v_1, v), (v, v')$  and  $(v', v_2)$  were the pairs of direct descendants in  $T$  which means  $v_2$  was the direct descendant of  $v_1$  by transitivity of this relation.



Lemma 11 implies that a terminal vertex being in scope of a pump in a collapsed derivation tree was also in scope of this pump in the original tree. This fact allows us to prove a weakened analogue of the Ogden lemma ([6]).

**Theorem 3 (Ogden lemma for 1-DCFGs).** *For any 1-DCFL  $L$  there is a number  $p$  such that for any word  $w \in L$  with at least  $p$  selected positions there is a representation  $w = x_0 y_1 u_1 z_1 x_1 y_2 u_2 z_2 x_2$  satisfying the following conditions:*

1. *For any  $m \in \mathbb{N}$  the word  $x_0 y_1^m u_1 z_1^m x_1 y_2^m u_2 z_2^m x_2$  belongs to  $L(G)$ .*
2. *There is at least one selected position in some of the words  $y_1, z_1, y_2, z_2$ .*

*Proof.* We set  $p$  equal to  $n$  from pumping lemma. It suffices to show that one of selected positions is in scope of some pump. We use induction on  $|w|$ , note that this length is at least  $n$ . There is a presentation  $w = x'_0 y'_1 u'_1 z'_1 x'_1 y'_2 u'_2 z'_2 x'_2$  such that the word  $x'_0 u'_1 x'_1 u'_2 x'_2$  is also in  $L$ . If the removed words contained a labeled position, the lemma is proved. Otherwise the word  $w' = x'_0 u'_1 x'_1 u'_2 x'_2$  contains the same number of labeled positions and we can apply the induction hypothesis to its derivation tree  $T'$ , which is obtained from  $T$  by collapsing. Then one of selected positions is in scope of some pump in  $T'$ , which implies by Lemma 11 it was in scope of a pump in  $T$  already. The theorem is proved.

## 5 Examples of non 1-DCFLs

In this section we use the established theoretical results to give some examples of non-1-DCFLs. To address this question we need to investigate more thoroughly the properties of constituents of displacement context-free grammars. A constituent is the fragment of the word derived from a node of derivation tree. In the context-free case every constituent is a continuous subword, hence it can be described by two numbers: the position of its first symbol and the position of its last symbol plus one (we add one to deal with empty constituents). Recall that context-free constituents must be correctly embedded which means they either do not intersect or one constituent is the part of the another.

The situation is a bit more complex in the case of DCFGs. However, the results of [10] provide analogous geometrical intuition. The constituents of rank 1 are the words of the form  $w_1 l w_2$ , where  $w_1, w_2$  are continuous subwords of the derived word  $w$ . Then a constituent of rank 1 is characterized by four indexes  $i_1 \leq j_1 \leq i_2 \leq j_2$  of the border of its subwords. We identify a constituent with a tuple of its characterizing indexes in the ascending order. The proofs of the statements below are carried out to the Appendix.

**Lemma 12.** *One of the possibilities below hold without loss of generality for any pair of constituents  $(i_1, j_1, i_2, j_2)$  and  $(i'_1, j'_1, i'_2, j'_2)$ :*

1.  $j_2 \leq i'_1$ ,
2.  $j_1 \leq i'_1 \leq j'_2 \leq i_2$ ,
3.  $i_1 \leq i'_1 \leq j'_2 \leq j_1$  or  $i_2 \leq i'_1 \leq j'_2 \leq j_2$ ,
4.  $i_1 \leq i'_1 \leq j'_1 \leq j_1 \leq i_2 \leq i'_2 \leq j'_2 \leq j_2$ .

Since every pump is just a pair of properly embedded constituents labeled by the same nonterminal, Lemma 12 helps to specify the mutual positions of different 4-pumps. The scope of the pump contains exactly the positions which are in the top constituent but not in the bottom, so every 4-pump is described by eight indexes  $i_1 \leq j_1 \leq k_1 \leq l_1 \leq i_2 \leq j_2 \leq k_2 \leq l_2$ , such that  $(i_1, l_1, i_2, l_2)$  is the tuple of indexes of its top constituent and  $(j_1, k_1, j_2, k_2)$  — of the bottom.

**Lemma 13.** *One of the possibilities below hold without loss of generality for any pair of 4-pumps  $(i_1, j_1, k_1, l_1, i_2, j_2, k_2, l_2)$  and  $(i'_1, j'_1, k'_1, l'_1, i'_2, j'_2, k'_2, l'_2)$ :*

1.  $l_2 \leq i'_1$ ,
2.  $i_1 \leq i'_1 \leq l'_2 \leq j_1$  or  $k_2 \leq i'_1 \leq l'_2 \leq l_2$ ,
3.  $i_1 \leq i'_1 \leq j'_1 \leq j_1 \leq k_1 \leq k'_1 \leq l'_1 \leq l_1 \leq i_2 \leq i'_2 \leq j'_2 \leq j_2 \leq k_2 \leq k'_2 \leq l'_2 \leq l_2$ ,
4.  $i_1 \leq i'_1 \leq j'_1 \leq k'_1 \leq j_1 \leq k_1 \leq l'_1 \leq l_1 \leq i_2 \leq i'_2 \leq j'_2 \leq k'_2 \leq j_2 \leq k_2 \leq l'_2 \leq l_2$ ,
5.  $i_1 \leq i'_1 \leq j_1 \leq k_1 \leq j'_1 \leq k'_1 \leq l'_1 \leq l_1 \leq i_2 \leq i'_2 \leq j'_2 \leq k'_2 \leq j_2 \leq k_2 \leq l'_2 \leq l_2$ ,
6.  $i_1 \leq i'_1 \leq j_1 \leq j'_1 \leq k'_1 \leq k_1 \leq l'_1 \leq l_1 \leq i_2 \leq i'_2 \leq j'_2 \leq k'_2 \leq j_2 \leq k_2 \leq l'_2 \leq l_2$ ,
7.  $k_1 \leq i'_1 \leq l'_1 \leq l_1 \leq i_2 \leq i'_2 \leq l'_2 \leq j_2$ ,
8.  $i_1 \leq i'_1 \leq l'_1 \leq j_1 \leq k_2 \leq i'_2 \leq l'_2 \leq l_2$ ,
9.  $k_1 \leq i'_1 \leq l'_2 \leq l_1$  or  $i_2 \leq i'_1 \leq l'_2 \leq j_2$ ,
10.  $j_1 \leq i'_1 \leq l'_1 \leq k_1 \leq j_2 \leq i'_2 \leq l'_2 \leq k_2$ ,
11.  $j_1 \leq i'_1 \leq l'_2 \leq k_1$  or  $j_2 \leq i'_1 \leq l'_2 \leq k_2$ ,
12.  $l_1 \leq i'_2 \leq l'_2 \leq i_2$ .

Let  $\pi_1 = (i'_1, j'_1, k'_1, l'_1, i'_2, j'_2, k'_2, l'_2)$  and  $\pi_2 = (i_1, j_1, k_1, l_1, i_2, j_2, k_2, l_2)$  be two 4-pumps. We call a pair of  $\pi_1$  and  $\pi_2$  linear if  $l_2 \leq i'_1$  or  $l'_2 \leq i_1$ . We call  $\pi_1$  outer for the pump  $\pi_2$  if  $i_1 \leq i'_1 \leq l'_2 \leq l_2$ . Note that if a pair of 4-pumps is not linear, then one of its elements is the outer pump for another. We call  $\pi_1$  embracing for the second if  $l_1 \leq i'_1 \leq l'_2 \leq i_2$ .

**Corollary 1.** *Let  $(i_1, j_1, k_1, l_1, i_2, j_2, k_2, l_2)$  and  $(i'_1, j'_1, k'_1, l'_1, i'_2, j'_2, k'_2, l'_2)$  be 4-pumps such that one of the segments of the second pump is a proper subset of the segment  $[l_1; i_2]$ . Then either the second pump is outer for the first (which means  $i'_1 \leq i_1 \leq l_2 \leq l'_2$ ) or the first pump is embracing for the second.*

Lemma 13 allows us to give some examples of non 1-MCFLs. The first example is the language  $4MIX = \{w \in \{a, b, c, d\}^* \mid |w|_a = |w|_b = |w|_c = |w|_d\}$ .

**Theorem 4.** *The language 4MIX cannot be generated by any 1-MCFG.*

*Proof.* Since wMCFLs are closed under intersection with regular languages, it suffices to prove that the language  $4MIX \cap (a^+b^+c^+d^+)^2$  is not a 1-DCFL. Assume the contrary, let  $t$  be the number from Ogden's lemma applied to this language. Let the word  $w = a^{m_1}b^{m_2}c^{m_3}d^{m_4}a^{n_1}b^{n_2}c^{n_3}d^{n_4}$  satisfy the following conditions:

1.  $\min(m_j, n_j) \geq t$ ,
2.  $m_1 \geq (3M + 1)(M + t)$ , where  $M = \max(m_2, m_4, n_3)$ ,
3.  $m_4 \geq (n_1 + 1)(n_1 + t)$ .

Note that every 4-pump contains the equal number of  $a$ -s,  $b$ -s,  $c$ -s and  $d$ -s, and every continuous segment of it consists of identical symbols (we call such segments homogeneous). We enumerate the maximal continuous homogeneous subwords of  $w$  from 1 to 8. Then every 4-pump intersects with exactly 4 of such segments. We call a  $x_1 \dots x_l$ -pumping group a pump intersecting the segments with numbers  $x_1, \dots, x_l$  (and possibly some others).

We select  $3M + 1$  segments of length  $t$  in the first segment of the word  $w$  so, that any two segments are separated by not less than  $M$  symbols. By Ogden lemma each such segment intersects with some 4-pump. We want to prove that some of them intersects with 1368-pumping group. Indeed, any two points from different segments cannot belong to the same 17-group since in this case there is a continuous segment of at least  $M + 1$   $a$ -s in the pump, then the pump contains at least  $M + 1$   $c$ -s, which exceeds the length of the 7-th segment. By the same arguments there are at most  $M$  12-pumping groups and at most  $M$  14-pumping groups, therefore the number of 1-groups which are not 1368-groups is less than  $3M + 1$  which proves the existence of a 1368-pumping group.

By the same arguments there is at least one 4-group, which is not a 45-group. By corollary 1 applied to the 1368-pumping group it is either a 148-group or it is embraced by the 1368-pumping group. In the first case there are two  $d$  segments in the pump, in the second case it should be a 3456-group which contradicts our assumption. So we have reached a contradiction and the theorem is proved.

Our technique of embedding different 4-pumps also works in a more complex case. Consider the language  $MIX = \{w \in \{a, b, c\}^* \mid |w|_a = |w|_b = |w|_c\}$ . It is expected to be not a *DCFL* since it demonstrates an extreme degree of unprojectivity. It is proved in [4] that  $MIX$  is not a 2-wMCFL (and hence not a 1-DCFL). The proof extensively uses geometrical arguments and is therefore very difficult to be generalized for similar languages or wMCFGs of higher order. Our proof uses only the Ogden's lemma and is much shorter.

**Theorem 5.** *The  $MIX$  language is not a 1-DCFL.*

*Proof.* We use the same method as in the case of 4MIX language. Again, it suffices to prove that the language  $L = MIX \cap a^+b^+c^+b^+c^+a^+$  is not a 1-DCFL. Let  $t$  be the number from Ogden's lemma for  $L$ . Consider the word  $w = a^{m_1}b^{m_2}c^{m_3}b^{n_2}c^{n_3}a^{n_1}$  satisfying the following properties:

1.  $\min(m_j, n_j) \geq t$ ,
2.  $m_1 \geq (2M + 1)(M + t)$ , where  $M = \max(m_3, n_2)$ ,
3.  $n_1 \geq (2M + 1)(M + t)$ , where  $M = \max(m_3, n_2)$ ,
4.  $m_3 \geq (n_2 + 1)(n_2 + t)$ .

By the same arguments as in Theorem 4 we establish the existence of 125- and 256-pumping groups. Since they cannot form a linear pair of 4-pumps, one of them is an outer pump for another, which implies one of them is a 1256-pumping group. The condition  $m_3 \geq (n_2 + 1)(n_2 + t)$  implies the existence of a 23-pumping group. By Corollary 1 applied to the 1256-pumping group and the 23-group, the 23-group is actually a 1236-group since it contains some  $a$ -s.

The condition  $n_2 \geq t$  implies there is a 4-group, which is a 146-group again by Corollary 1. To be correctly embedded with the 1236-group it should be a 1246 group but there is no  $c$ -s in such group. Hence we reached the contradiction and the *MIX*-language cannot be generated by a 1-DCFG. The theorem is proved.

## 6 Conclusions and future work

We have proved a strong version of the pumping lemma and an analogue of Ogden’s lemma for the class of 1-DCFLs which is also the class of tree-adjoining languages. These statements allow us to prove that some languages, like the well-known *MIX*-language, do not belong to this family. We hope to adopt the proof for the case of semiblind three-counter language  $\{w \in \{a, b, c\}^* \mid |w|_a = |w|_b = |w|_c, \forall u \sqsubseteq w \mid u|_a \geq |u|_b \geq |u|_c\}$  to prove that a shuffle iteration of a one-word language may lie outside the family of 1-DCFLs. The author supposes that pumping lemma is valid for the class of  $k$ -DCFGs under natural replacement of 4-pumps with  $(2k+2)$ -pumps. We hope that the established results will help us to understand better the structure of well-nested MCFLs and, in particular, prove the Kanazawa conjecture, which states that *MIX* is not a well-nested MCFL.

## References

1. Aravind Joshi. Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? University of Pennsylvania, Moore School of Electrical Engineering, Department of Computer and Information Science, 1985, 206–250.
2. Aravind Joshi and Yves Schabes. Tree-adjoining grammars. In G. Rozenberg and A. Salomaa, eds., *Handbook of Formal Languages*, Vol. 3: Beyond Words, 69–123. Springer, Berlin, 1997.
3. Makoto Kanazawa. The pumping lemma for well-nested multiple context-free languages. *Developments in Language Theory, Lecture Notes in Computer Science*, 5583 (2009) 312–325.
4. Makoto Kanazawa, Sylvain Salvati. *MIX* is not a tree-adjoining language. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (2012), 666–674.
5. Glyn Morrill, Oriol Valentín and Mario Fadda. The displacement calculus. *Journal of Logic, Language and Information*. 20(1) (2011) 1–48.
6. William Ogden. A helpful result for proving inherent ambiguity. *Mathematical systems theory*. 2(3) (1968) 191–194.
7. Carl Pollard. Generalized phrase structure grammars, head grammars, and natural languages. Ph.D. thesis, Stanford University, Stanford, 1984.
8. Hiroyuki Seki, Takashi Matsumoto, Mamoru Fujii and Tadao Kasami. On multiple context-free grammars. *Theoretical Computer Science* 88(2) (1991) 191–229.
9. Alexey Sorokin. Normal forms for multiple context-free languages and displacement Lambek grammars. *Logical Foundations of Computer Science, Lecture Notes in Computer Science*, 7734 (2013), 319–334.

10. Alexey Sorokin. Monoid automata for displacement context-free languages. ESSLLI Student Session 2013 Preproceedings. (2013) 158–167. <http://loriweb.org/uploads/ESSLLIStuS2013.pdf#page=162>  
Extended version to appear in ESSLLI Student Session 12-13 Selected Papers, <http://arxiv.org/abs/1403.6060>
11. Aravind Joshi, K. Vijay-Shanker and David Weir. Tree adjoining and head wrapping. Proceedings of the 11th Conference on Computational Linguistics (1986), 202–207.

## A Term equivalence

**Lemma 14.** *For any  $k$ -essential term  $\alpha$  there is an equivalent  $k$ -correct term  $\alpha'$ .*

*Proof.* At first we prove that there is an equivalent term with no internal subterms of rank greater than  $k$ . Let  $K$  be the maximal rank of subterms in  $\alpha$ , a subterm occurrence is called heavy if the rank of the corresponding subterm equals  $K$ . We use induction on  $K$  and the number of heavy subterm occurrences.

Let  $(\beta, v)$  be an occurrence a subterm  $\beta$  of rank  $K$  with minimal depth among all such occurrences.  $v$  cannot be the root of  $tree(\alpha)$  since  $rk(\alpha) < K$  so let  $(\gamma, v_1)$  be the subterm occurrence corresponding to its parent in the syntactic tree. Then  $rk(\gamma) < rk(\beta)$  which implies that  $\gamma = \beta \odot_j \eta$  for some 0-ranked term  $\eta$ . We transform  $\gamma$  to an equivalent subterm  $\gamma'$  with less occurrences of heavy subterms. The transformation uses the equivalences from Lemma 1.

Consider the possible structure of the term  $\beta$ . First, let it has the form  $\beta = \beta_1 \cdot \beta_2$ . If  $rk(\beta_1) \leq j$  then the term  $\gamma' = (\beta_1 \odot_j \eta) \cdot \beta_2$  is equivalent to  $\gamma$  and has less occurrences of subterms of rank  $K$  because we have removed the occurrence of subterm  $\beta$  and haven't add any other heavy subterms. In case  $j > rk(\beta)$  the term  $\gamma' = \beta_1 \cdot (\beta_2 \odot_{j-rk(\beta_1)} \eta)$  does the same job.

Now let  $\beta$  has the form  $\beta = \beta_1 \odot_l \beta_2$ . If  $j < l$  then we define  $\gamma' = (\beta_1 \odot_j \eta) \odot_{l-1} \beta_2$ . In case  $l \leq j < l + rk(\beta_2)$  we set  $\gamma' = \beta_1 \odot_l (\beta_2 \odot_{j-l+l} \eta)$  and in case  $j \geq l + rk(\beta)$  we define  $\gamma' = (\beta_1 \odot_{l+rk(\beta_2)-1} \eta) \odot_j \beta_2$ . In all the cases  $\gamma'$  is equivalent to  $\gamma$  by lemma 1 and has fewer occurrences of heavy subterms.

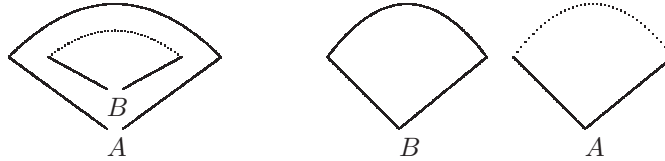
Since  $\gamma$  was a subterm of  $\alpha$ , there is a context  $C$  such that  $\alpha = C[\gamma]$ . Then the term  $\alpha'' = C[\gamma']$  is equivalent to  $\alpha$  and has fewer occurrences of heavy subterms. We can apply the induction hypothesis to  $\alpha''$  and obtain the required term  $\alpha$ .

In case  $k \geq 1$  the proof is completed since the rank of atomic subterms cannot be greater than 1, but we need a minor complication if  $k = 0$ . In this case some leaves of  $tree(\alpha')$  might be labeled by 1. However they all occur in subterms of the form  $1 \odot_1 \beta$  for some  $\beta$  of rank 0 since  $\alpha$  has no internal subterms of positive rank. If we replace all subterms of the form  $1 \odot_1 \beta$  by the corresponding term  $\beta$ , we obtain an equivalent 0-correct term. The lemma is proved.

## B Constituents in displacements context-free grammars

This section we discuss the geometrical interpretation of constituents in displacement context-free grammars. A constituent is a (possibly discontinuous)

fragment of a word derived from a node of its derivation tree. The nonterminal label of this node is the label of the constituent. In the basic context-free case the constituents are just continuous subwords, so every constituent is completely defined by two indexes  $i, j$ : the position of its first symbol and the position of its last symbol plus one (we add one to deal with empty constituents). Different constituents should satisfy the embedding conditions: either one of them is inside the other ( $[i; j] \subseteq [i'; j']$  or  $[i'; j'] \subseteq [i; j]$  in terms of indexes), or they do not have common internal points ( $[i; j] \cap [i'; j']$  is one of the sets  $\emptyset, \{i\}, \{j\}$ ). Mutual positions of different constituents are shown on the picture below.



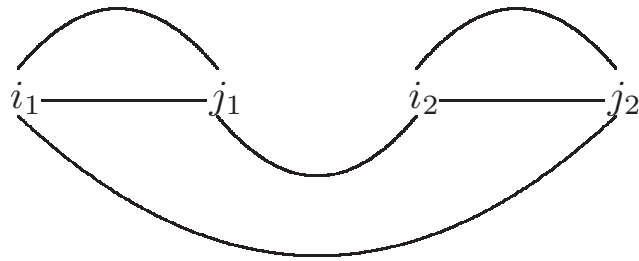
Let us now inspect the constituent structure of 1-DCFGs. In the case of these grammars every constituent is either a continuous subword, if its label is of rank 0, or a word of the form  $w_1 l w_2$  where  $w_1$  and  $w_2$  are continuous segments of the derived word  $w$ , if the label is of rank 1. We focus our attention on the latter case because nothing has changed from the context-free case for the constituents of rank 0. Then the first continuous part of the constituents is described by indexes  $i_1, j_1$  and the second part by indexes  $i_2, j_2$ . Therefore every constituent of rank 1 corresponds to a tuple  $(i_1, j_1, i_2, j_2)$  of its indexes taken in the ascending order. We will not distinguish constituents and their index tuple in the further.

The following lemma about mutual positions of different constituents was proved in [10] in a more general case.

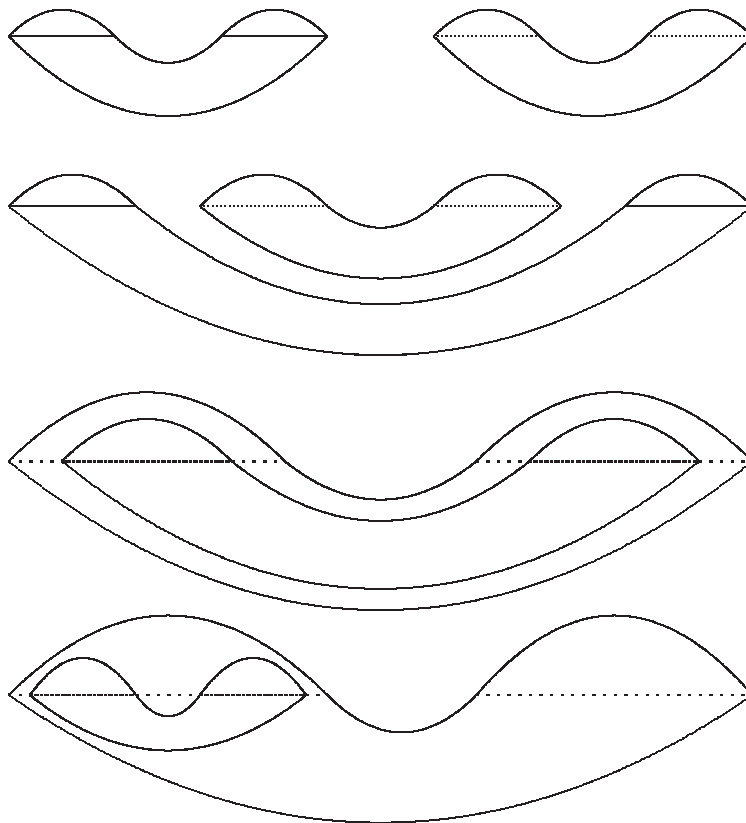
**Lemma 15.** *One of the possibilities below hold without loss of generality for any pair of constituents  $(i_1, j_1, i_2, j_2)$  and  $(i'_1, j'_1, i'_2, j'_2)$ :*

1.  $j_2 \leq i'_1$ ,
2.  $j_1 \leq i'_1 \leq j'_2 \leq i_2$ ,
3.  $i_1 \leq i'_1 \leq j'_2 \leq j_1$  or  $i_2 \leq i'_1 \leq j'_2 \leq j_2$ ,
4.  $i_1 \leq i'_1 \leq j'_1 \leq j_1 \leq i_2 \leq i'_2 \leq j'_2 \leq j_2$ .

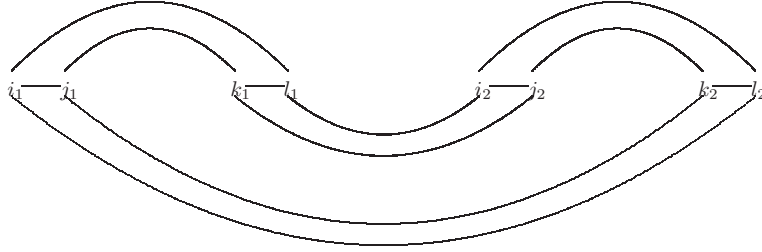
The statement of the lemma above has a nice geometrical interpretation. We associate with every constituent  $(i_1, j_1, i_2, j_2)$  of rank 1 the following curve (the constituents themselves are marked by horizontal lines):



The remarkable property of this interpretation is that if we write a derived word of the abscissa axis, enumerate the positions in it and draw the curves corresponding to all its constituents, then these curves must not intersect except the limit points. On the picture below we show all principal variants of different constituents location (solid and dash horizontal lines mark the constituent itself).



Provided geometrical interpretation is very helpful in our main task: studying mutual positions of different pumps. Indeed, every pump is defined by its top and bottom nodes, which carry the same nonterminal labels and are connected by the path of nodes of the same rank. Since every node of the derivation tree corresponds to a constituent, then a pump is matched with a pair of embedded constituents with the same label. As earlier, we concentrate on the 4-pumps which correspond to a pair of constituents of rank 1. Then every pump can be defined by 8 numbers  $i_1 \leq j_1 \leq k_1 \leq l_1 \leq i_2 \leq j_2 \leq k_2 \leq l_2$  such that  $(i_1, l_1, i_2, l_2)$  are the indexes of its top constituent and  $(j_1, k_1, j_2, k_2)$  — of the bottom. We call the segments  $[i_1; j_1]$ ,  $[k_1; l_1]$ ,  $[i_2; j_2]$ ,  $[k_2; l_2]$  the segments of the pump and identify a pump with the ascending tuple of its indexes. Below we illustrate how two constituents of rank 1 with the same label form a 4-pump:



Since the curves on the picture are the bounding curves for the constituents forming the pump, the curves corresponding to different pumps must not intersect anywhere except the abscissa axis. The following lemma interprets the geometrical conditions on correct embedding in terms of pump segments:

**Lemma 16.** *One of the possibilities below hold without loss of generality for any pair of 4-pumps  $(i_1, j_1, k_1, l_1, i_2, j_2, k_2, l_2)$  and  $(i'_1, j'_1, k'_1, l'_1, i'_2, j'_2, k'_2, l'_2)$ :*

1.  $l_2 \leq i'_1$ ,
2.  $i_1 \leq i'_1 \leq l'_2 \leq j_1$  or  $k_2 \leq i'_1 \leq l'_2 \leq l_2$ ,
3.  $i_1 \leq i'_1 \leq j'_1 \leq j_1 \leq k_1 \leq k'_1 \leq l'_1 \leq l_1 \leq i_2 \leq i'_2 \leq j'_2 \leq j_2 \leq k_2 \leq k'_2 \leq l'_2 \leq l_2$ ,
4.  $i_1 \leq i'_1 \leq j'_1 \leq k'_1 \leq j_1 \leq k_1 \leq l'_1 \leq l_1 \leq i_2 \leq i'_2 \leq j'_2 \leq k'_2 \leq j_2 \leq k_2 \leq l'_2 \leq l_2$ ,
5.  $i_1 \leq i'_1 \leq j_1 \leq k_1 \leq j'_1 \leq k'_1 \leq l'_1 \leq l_1 \leq i_2 \leq i'_2 \leq j'_2 \leq k'_2 \leq j_2 \leq k_2 \leq l'_2 \leq l_2$ ,
6.  $i_1 \leq i'_1 \leq j_1 \leq j'_1 \leq k'_1 \leq k_1 \leq l'_1 \leq l_1 \leq i_2 \leq i'_2 \leq j'_2 \leq k'_2 \leq k_2 \leq l'_2 \leq l_2$ ,
7.  $k_1 \leq i'_1 \leq l'_1 \leq l_1 \leq i_2 \leq i'_2 \leq l'_2 \leq j_2$ ,
8.  $i_1 \leq i'_1 \leq l'_1 \leq j_1 \leq k_2 \leq i'_2 \leq l'_2 \leq l_2$ ,
9.  $k_1 \leq i'_1 \leq l'_2 \leq l_1$  or  $i_2 \leq i'_1 \leq l'_2 \leq j_2$ ,
10.  $j_1 \leq i'_1 \leq l'_1 \leq k_1 \leq j_2 \leq i'_2 \leq l'_2 \leq k_2$ ,
11.  $j_1 \leq i'_1 \leq l'_2 \leq k_1$  or  $j_2 \leq i'_1 \leq l'_2 \leq k_2$ ,
12.  $l_1 \leq i'_2 \leq l'_2 \leq i_2$ .

*Proof.* The present lemma may be proved by geometrical arguments only, however, we derive it formally from Lemma 15. We call a pair of 4-pumps linear if



$l_2 \leq i'_1$  or  $l'_2 \leq i_1$ . We call the pump  $(i_1, j_1, k_1, l_1, i_2, j_2, k_2, l_2)$  outer for the pump  $(i'_1, j'_1, k'_1, l'_1, i'_2, j'_2, k'_2, l'_2)$  if the condition  $i_1 \leq i'_1 \leq l'_2 \leq l_2$  holds. Note that if two pumps do not form a linear pair, then one of them is an outer for the other.

We denote  $\pi = (i_1, j_1, k_1, l_1, i_2, j_2, k_2, l_2)$  and  $\pi' = (i'_1, j'_1, k'_1, l'_1, i'_2, j'_2, k'_2, l'_2)$  to shorten the notation. If the pair of  $\pi$  and  $\pi'$  is linear then up to renaming the pumps the first alternative of the lemma holds. Otherwise one of the pumps is the outer for another, let  $\pi$  be such a pump. So  $i_1 \leq i'_1 \leq l'_2 \leq l_2$ . Consider the constituents  $(i_1, l_1, i_2, l_2)$ ,  $(j_1, k_1, j_2, k_2)$ ,  $(i'_1, l'_1, i'_2, l'_2)$ ,  $(j'_1, k'_1, j'_2, k'_2)$ , each of them bounds a region on the plane. By the geometric interpretation of Lemma 15 for any pair of such regions there are only two possibilities either the elements of the pair do not intersect or the smaller constituent is inside the bigger.

Consider at first the case when the regions of the constituents  $(i_1, l_1, i_2, l_2)$  and  $(i'_1, l'_1, i'_2, l'_2)$  do not intersect. Since the segment  $[i'_1; l'_2]$  is a subset of the segment  $[i_1; l_2]$  it is possible only when  $l_1 \leq i'_1 \leq l'_2 \leq i_2$ , which is one of the alternatives provided by the present lemma.

In the other case the region corresponding to the constituent  $(i_1, l_1, i_2, l_2)$  contains all the other regions. We consider different variants of embedding of the constituents. If constituents  $(j_1, k_1, j_2, k_2)$  and  $(i'_1, l'_1, i'_2, l'_2)$  do not intersect, then either  $l'_2 \leq j_1$ ,  $k_2 \leq i'_1$ ,  $l'_1 \leq j_1 \leq k_2 \leq i'_2$  or  $k_1 \leq i'_1 \leq l'_2 \leq j_2$ . In the first case  $i_1 \leq i'_1 \leq l'_2 \leq j_1$ , symmetrically in the second  $k_2 \leq i'_1 \leq l'_2 \leq l_2$ , and in the third case  $i_1 \leq i'_1 \leq l'_1 \leq j_1 \leq k_2 \leq i'_2 \leq l'_2 \leq l_2$  which all satisfy the requirements of the present lemma. Consider the last subcase  $k_1 \leq i'_1 \leq l'_2 \leq j_2$ , then applying the Lemma 15 to the constituents  $(i_1, l_1, i_2, l_2)$  and  $(i'_1, l'_1, i'_2, l'_2)$  we obtain that either  $l'_2 \leq l_1$ ,  $i_2 \leq i'_1$  or  $l'_1 \leq l_1 \leq i_2 \leq i'_2$ . Taking into account all the inequalities, we obtain that  $k_1 \leq i'_1 \leq l'_2 \leq l_1$  or  $i_2 \leq i'_1 \leq l'_2 \leq j_2$  or  $k_1 \leq i'_1 \leq l'_1 \leq l_1 \leq i_2 \leq i'_2 \leq l'_2 \leq j_2$ , which is allowed by the lemma.

Now consider the case when the region of the constituent  $(i'_1, l'_1, i'_2, l'_2)$  is inside the region of  $(j_1, k_1, j_2, k_2)$ . It means that one of the following possibilities hold:  $j_1 \leq i'_1 \leq k'_2 \leq k_1$ ,  $j_2 \leq i'_1 \leq l'_2 \leq k_2$  or  $i'_1 \leq j_1 \leq k_1 \leq l'_1 \leq i'_2 \leq j_1 \leq k_2 \leq l'_2$ . All these variants satisfy the requirements of the Lemma.

So it remains to inspect the case when the region of constituent  $(i'_1, l'_1, i'_2, l'_2)$  includes the region of  $(j_1, k_1, j_2, k_2)$ . Then  $i_1 \leq i'_1 \leq j_1 \leq k_1 \leq l'_1 \leq l_1 \leq i_2 \leq i'_2 \leq j_2 \leq k_2 \leq l'_2 \leq l_2$  and we should consider the mutual positions of the regions of constituents  $(j_1, k_1, j_2, k_2)$  and  $(j'_1, k'_1, j'_2, k'_2)$ . This leads us to the following variants:

$$\begin{aligned} i'_1 &\leq j'_1 \leq k'_1 \leq j_1 \leq k_1 \leq l'_1 \leq i'_2 \leq j_2 \leq k_2 \leq j'_2 \leq k'_2 \leq l'_2, \\ i'_1 &\leq j_1 \leq j'_1 \leq k'_1 \leq k_1 \leq l'_1 \leq i'_2 \leq j_2 \leq j'_2 \leq k'_2 \leq k_2 \leq l'_2, \\ i'_1 &\leq j_1 \leq k_1 \leq j'_1 \leq k'_1 \leq l'_1 \leq i'_2 \leq j'_2 \leq k'_2 \leq j_2 \leq k_2 \leq l'_2, \\ i'_1 &\leq j'_1 \leq j_1 \leq k_1 \leq k'_1 \leq l'_1 \leq i'_2 \leq j'_2 \leq j_2 \leq k_2 \leq k'_2 \leq l'_2, \end{aligned}$$

But all such variants are allowed the lemma conclusion. All the cases have been verified and the lemma is proved.